CLASSIFICATION VERSUS INFERENCE LEARNING CONTRASTED WITH REAL-WORLD CATEGORIES

BY

ERIN L. JONES

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2010

Urbana, IL

Advisor:

Professor Brian H. Ross

ABSTRACT

A number of studies have contrasted classification and inference, using a variety of stimuli and tests, but in all cases the difference between these tasks could be attributed to either methodological differences or an inherent difference between the tasks. The inherent difference explanation argues that classification and inference learners use different strategies during learning, which reflect the goals of the tasks. Inference learners focus more on what each category is like, while classification learners focus on finding the information that best predicts category membership. These differences during learning lead to performance differences on later tests of category knowledge. In two experiments, using real-world categories and controlling for methodological differences, inference learners learned more about what each category was like than classification learners. These results suggest that there is an inherent difference between classifying an item and inferring a feature that cannot be explained by methodological differences between the tasks.

ACKNOWLEDGEMENTS

I would like to thank Brian Ross, my advisor and collaborator on this project, for taking the time to train and prepare me to produce and write works such as this one. I would also like to thank Kara Federmeier for providing insightful comments and feedback, and Eric Taylor and Wookyoung Jung for their comments on an earlier draft of this paper.

I would also like to acknowledge Chris Wahlheim and Larry Jacoby, who generously provided the bird pictures from their Bird Learning project, and Robert Molitor for collecting and organizing the data.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Categories underlie a variety cognitive tasks, including problem solving, classification, prediction, explanations, and inferences. Knowing the category an object belongs to enables access to categorical knowledge acquired from past experiences. We use category knowledge when retrieving the formula for solving permutation problems, when classifying an approaching person as a soldier, or when deciding whether to flee from an approaching Labrador retriever. One can argue that the importance of categories is not in knowing that the object in front of you is a chair or a table, but rather in using the category knowledge associated with the object to make inferences about it, such as whether to sit on it or place a glass on it (Anderson, 1991; Murphy, 2002). We learn categories in many ways and it is likely that differences in processing during learning may lead to differences in category knowledge (Markman & Ross, 2003). Despite this variety, the majority of research on category learning focuses on classification learning and largely ignores other tasks.

Recently, a number of studies have contrasted two major means of category learning: classification, where the learner determines what category an item belongs to, and inference, where the learner predicts a feature of a classified item (e.g. Yamauchi & Markman, 1998; Anderson, Ross, & Chin-Parker, 2002; Chin-Parker & Ross, 2004). The importance of this comparison is that classification and inference are two of the main functions of categories. The force of the comparison is that these tasks are formally equivalent if category label is treated as just another feature of the category (Anderson, 1991; Yamauchi & Markman, 1998). Thus, any differences between classification and inference must be due to the inherent nature of predicting either a category label or feature.

These studies generally use a family resemblance category structure (Rosch & Mervis, 1975) with binary-valued features. Table 1 shows an example family resemblance category structure with 4 binary dimensions. For each category there is a value for each dimension that is prototypical, or common across the category members, and atypical for the other category's members. For instance, if striped wings are prototypical for category A, and an exemplar has striped wings, this would be signified by a 1 in the first position. To capture variation across category members, three out of four of an exemplar's dimensions are its category's prototypical values. The fourth dimension, known as the *exception feature*, is the other category's prototypical value.

In the classification learning paradigm, participants are presented with an exemplar from one of two categories, asked to classify the exemplar, and provided with feedback. For example, they might be shown some fictional creature and asked if it is a Deeger or Koozle. In the inference learning paradigm, participants are presented with an exemplar from one of two categories along with its category label but without one of its features, asked to provide the missing feature's value, and provided with feedback. Thus, here the creature might be presented with a Deeger label but without its tail and the subject is given the two tail choices to choose from. In both learning conditions, the learning trials continue until the participant reaches a predetermined learning criterion or a fixed number of trials set by the experimenter. The critical difference between the two tasks is the type of information queried during learning – category label or feature.

Research has shown that classification and inference learners perform differently on a variety of category tasks (Yamauchi & Markman, 1998; Anderson, Ross, & Chin-Parker, 2002; Chin-Parker & Ross, 2004). The classification learners focus principally on the diagnostic features, those that distinguish the categories, whereas inference learners focus on features that are common to many category members, even if they are not diagnostic (Chin-Parker & Ross, 2004).

The critical issue is whether the observed performance differences reflect some inherent difference between classification and inference learning or between the ways these tasks are carried out in an experimental setting. The inherent-difference explanation for the performance differences is that classification and inference encourage learners to adopt learning strategies reflecting the task's goals, and consequently, learners develop category representations that are influenced by how they learned (Markman & Ross, 2003). In classification, the goal is to find the feature or features that provide the most diagnostic information so that given an item the learner can successfully tell what category it is in. In inference, the goal is to learn the common features (and correlations; Chin-Parker & Ross, 2002) of each category's exemplars, so that given an item and its category, the learner can successfully infer a not-presented feature. In most real-world categories, many of the features are interrelated, so the learner may be more likely to consider the other features when making an inference. So, although, classification and inference can be viewed as formally equivalent *if* a category label is considered as a feature, there are theoretical and empirical reasons to believe that the category label is not just a feature and that the two tasks differ in important ways (e.g., Gelman & Heyman, 1999; Markman & Ross, 2003).

An alternative explanation is that there is no inherent difference, but that classification and inference learners perform differently because of how these tasks are implemented in an experimental setting (e.g., Johansen & Kruschke, 2005). While many efforts have been made in the past to formally equate classification and inference, there are still problems with the comparison that could potentially explain these performance differences without requiring one to believe the tasks lead to inherent differences.

We can identify three main differences between how the two tasks have commonly been conducted that might lead to performance differences even without inherent differences: number of queried dimensions, exception features, and the type of queried feature. First, and most notably, in all of the previous work comparing the two tasks, inference learners have been asked about multiple features while classification learners have been only asked to provide the category label throughout learning. For example, on some trials the inference learners might be asked to choose which tail the presented Deeger has and on other trials what legs a presented Deeger has. Inference learners could be learning more about what each category is like simply because they are explicitly asked about more of the category's features.

Although this difference might be an important one in real-world situations, it makes these comparisons questionable. Anderson, et al. (2002) found that inference learners who only inferred 2 out of 4 of the features still learned more about the other features than did classification learners. However, Rehder, Colner, and Hoffman (2009) proposed that inference learners' attention to these non-queried features is not some critical part of inference learning, but rather because in such an experiment the inference learners anticipate being asked about these features on later learning trials (whereas classification learners know that they will only be asked about the label). Rehder, et al. showed that when participants are told that they will only be asked about some of the features, their looking time at the non-queried features decreases drastically, though the features were quite separated to allow easy eye-tracking measurements. The separateness of the features within an item makes the arbitrary nature of artificial categories' features more salient, and could make differences between classification and inference more difficult to see. One reason why classification and inference might differ is because features are meaningfully related to each other and not arbitrary, while a category label is arbitrary with respect to the features of the category. When learning about a feature, it is useful to look at how it fits with other features within the category. A real-world domain would be better suited for examining this issue because participants will be more likely to assume that the features are meaningful and related (e.g. if a Finch eats insects, one would assume that the Finch's beak is shaped for optimal insect consumption). So, the question still remains as to whether the greater learning of the common features with inference learning is due to inherent differences in the task or due to a difference in the number of features queried.

Second, in previous work using the inference paradigm and a family resemblance category structure as in Table 1, inference learners were never asked to infer an exception feature. For instance, they never encountered a trial where they must infer the third dimension for the exemplar 1101. Consequently, for inference learners, the category label is perfectly predictive of the feature value to be inferred, but such a relationship is not true for classification learners (no single feature perfectly predicts the category label across all trials). This difference between classification and inference could lead inference learners to focus more on the prototypical values of the category compared to classification learners, who are more aware of the exception features during learning.

Third, classification learners are always asked to provide a verbal label while inference learners are often asked to choose a visual feature. Learning about categories often includes learning about both visual information (what it looks like) and non-visual information (e.g. what it eats, its ferocity). Being asked about visual information could cause learners to pay more attention to other visual information than if they were asked about verbal information. Although one could use verbal stimuli to avoid this difference, many of the tasks showing inference-classification differences have used visual stimuli.

CHAPTER 2

THE CURRENT EXPERIMENTS

The goal of the current experiments is to examine how classification and inference learning affect subsequent category performance without these methodological differences and in a real world domain.

To equate classification and inference learning, we have addressed the three methodological problems (number of queried dimensions, exception features, and type of queried feature). First, we equated the amount of information queried during learning—inference learners only inferred one feature throughout learning while classification learners only inferred the category label. Second, to address the exception feature problem, we used a category structure that had 2 defining features (i.e. features that were 100% predictive of category membership and thus perfectly correlated with the category label), one of which was chosen as the queried feature for the inference learner. The inference learner always received one defining feature, the category label and the visual stimulus, and the task was to infer the other defining feature. The classification learner was provided with both defining features and a visual stimulus on each trial and asked to infer the category label. Third, the defining features were verbal predicates, so inference and classification learners were both asked questions about verbal information. Thus both groups received the visual stimulus and two pieces of verbal information (either the two verbal predicates or one of them with the category label) and had to answer with the third piece of verbal information (the category label or the other verbal predicate).

One problem with making everything equivalent across the two learning tasks when using the simple, artificial categories typically used is if only one feature is queried throughout, participants may turn the inference task into a classification task (e.g. by saying "Those are the triangle-eyed bugs and those are the square-eyed bugs"). If participants re-structure the task to make the inference task into a classification task, it is unlikely that any performance differences would remain. We sought to reduce the chance of this occurring by using more complex, real world bird categories (from Wahlheim, Jacoby, & Coane, submitted) for which simple feature descriptions and the category structures would be less obvious, plus would engage participants who tried to learn what the categories were like. An additional benefit to using real world categories is that the labels are real bird names, whereas the labels for experimenter created artificial categories are made-up and arbitrarily assigned. Participants may be more likely to treat these labels as meaningful, knowing that they correspond to instances in the real world. To our knowledge, there has only been one other study (Sakamoto & Love, 2006) that used complex, real world categories when comparing classification and inference (though they did not equate the tasks in the other ways done here).

We expected classification learners would focus on whatever information was most diagnostic for classification (the verbal predicates), whereas the inference learners would attend to information relevant for inferring the verbal predicate (the category label and/or other verbal predicate) but also attend to what category members are like, including the visual stimulus. Inference learners have reason to believe that the feature they are inferring relates to other features of the birds whereas classification learners have no reason to assume that the same is true of a category label. For instance, the shape of the bird's beak could be a useful cue for determining what a bird eats, but there is usually no meaningful relationship between beak shape and the bird's name.

To test for a difference between classification and inference learners, we used a novel classification test with only visual information (the verbal predicates were omitted). Neither learning task required participants to learn about the visual properties of the birds, so performance above chance on this test would indicate that learners attended to and acquired knowledge beyond what was necessary during learning. We hypothesized that if inference learning leads people to learn what each category is like, not just what is necessary for making the inference, then inference learners will perform better on tests that probe knowledge of what they have learned about the visual characteristics of the bird categories, as evidenced by performance on a novel classification test (the usual measure of category knowledge from classification learning).

CHAPTER 3

EXPERIMENT 1

Experiment 1 investigated whether classification and inference learners develop different category representations for complex real-world categories. All participants learned about 6 different categories of birds (see Figure 1). During learning, participants were provided with 2 verbal predicates and a picture of a bird from one of the categories. Inference learners were shown one defining feature verbal predicate, the bird's picture, and the category label and were asked to provide the value of the other defining feature on each trial. Classification learners were shown both defining feature verbal predicates and the bird's picture and were asked to provide the category label on each trial. Both groups received feedback after their responses and learned for the same amount of time. Thus, during learning, both groups saw two pieces of perfectly correlated verbal information (either one verbal predicate and the category label or the two verbal predicates) and a picture, and were asked to respond with the other perfectly correlated verbal information (either the other verbal predicate or the category label).

At test, we examined whether the two groups had important differences in categorical information. The methodological changes from the usual contrast (in particular, the single queried feature and elimination of exception features) led us to focus on what other information about the categories had been learned besides the defining features. We chose the most common means of assessing category knowledge—classification of novel members of the categories (i.e., ones not seen during learning). For this test, the verbal information was omitted from the trials and learners had to use what they had learned about the visual properties of each bird category to make classification decisions. Performing well on this task requires participants to attend to information not necessary for the learning task (the bird pictures) and generalize across previously seen exemplars. We predicted that inference learners would classify novel members of the bird categories better than classification learners, even though it is a *classification* test, because they will have attended more to what category members are like, including information not necessary for the learning task. That is, even though the category label and defining verbal features are all that need to be looked at to do the classification or inference tasks, we expected that the inference learners would tend to examine the bird pictures more and learn more about the different families of birds.

We included two additional tests of participants' knowledge. One, a recognition test with classification (on each trial determine whether the presented item was seen earlier in the experiment or not and then decide which category it belongs to), to determine if the inherent difference between classification and inference is partially due to difference in memory for the specific items. For example, perhaps both groups examine the pictures equally well, but the classification learners focus more on the exact exemplars shown during study. If both types of learners attend to the visual information, but classification learners encode specific exemplars while inference learners generalize across exemplars, then classification learners might show higher recognition memory performance. Two, we included classification and inference tests of the verbal predicates (one at a time, without the bird pictures) to determine whether the verbal information was learned equally well in both conditions. However, the main issue is how the two groups perform on the novel classification test -- classifying items they have not seen before when no verbal information is presented. If inference learners attend more to the visual information to learn about what each category is like beyond what is necessary during learning, then inference learners should show higher novel classification performance.

Method

Design

Participants were randomly assigned to one of two between-subjects conditions: classification and inference learning. Assignment of verbal feature values to category as well as answer groupings[1] were counterbalanced across participants, resulting in 8 counterbalancing groups.

Participants

Participants were 24 undergraduate students from the University of Illinois who participated for course credit.

Materials

The materials were color images of birds from various families in the Passeriformes order acquired from http://www.whatbird.com, a bird identification website and compiled by Wahlheim, et al. (submitted). The six bird families were Finches, Flycatchers, Sparrows, Swallows, Thrushes, and Warblers. There were 15 exemplars used throughout the experiment for each of the 6 families. 6 exemplars from each family were used in the study phase, 6 exemplars from each family were used in the novel classification test, and 3 additional exemplars from each family were used for the recognition test. The pictures during study had the birds in a context (e.g., on a post, in a tree) to make clear the task involved learning real world categories. However, the test pictures all showed birds with just a white background to ensure the context was not influencing the judgments. Assignment of exemplar to each phase of the experiment was

[1] During learning only three answers appeared at the same time to make the task easier. The same three answers always appeared together and the answers in each group were counterbalanced across participants.

determined based on previous test performance data. Exemplars were chosen in order to keep the groups of families equated for difficulty.

The verbal predicates were the nest material each bird uses and the food each bird eats. There were six values for each of these two attributes. For nest material, the values were grass, leaves, moss, mud, silk, and sticks. Each verbal predicate had the same structure: This bird's nest is made of _______." For the food each bird eats, the values were berries, insects, nuts, plants, seeds, and worms. The structure of this verbal predicate was: This bird eats _______." The values were chosen based on what real birds use to make their nests and what they actually eat, but assignment of value to bird family was counterbalanced across subjects, so did not actually reflect what each of these bird families build their nests with and what they eat. Each verbal predicate was perfectly predictive of category membership.

Procedure

After giving their informed consent, participants rated how much knowledge they had about each of the six bird families on a scale of 1-5, to ensure that none of the participants already had knowledge about the birds. Next, participants were told that they would be learning about different bird families and would be tested on their knowledge of the families later.

First, participants completed a learning phase. For both the classification and inference learning conditions, there were 2 blocks of 36 trials. There were 6 exemplars from each of the 6 bird families, presented in a random order for each of the two blocks.

For the classification learning condition, participants were shown a picture along with the two verbal predicates (what the bird ate as well as the nest material it used). Figure 2 shows an example learning trial. These three pieces of information always appeared on the left half of the screen. During the learning phase, participants were presented with these pieces of information in the following order (with each piece of information remaining on the screen until the end of the trial): verbal Predicate 1, bird picture, and verbal predicate 2. For each new piece of information, there was a 3 s delay until the next piece of information was shown. The two verbal predicates were equally likely to be displayed as verbal predicate 1 or verbal predicate 2. Additionally, the verbal predicates were equally likely to be displayed above or below the bird picture. After all of the information was on the screen, the question "What kind of bird is this?" and three possible answers were displayed on the right side of the screen. Even though there were 6 bird families, only three possible answers were displayed to make the task easier. We grouped bird families into two sets of three and their category labels were shown in these same sets throughout the learning phase. Participants clicked on their choice using the mouse and there was no time limit to make a selection. Next, participants were given feedback (3s) as to whether they were correct or incorrect and the statement "This is a [Correct bird family]" was displayed.

For the inference learning condition, the trial was the same as the classification learning condition, but instead of a second verbal predicate, participants saw the statement "This is a [Correct bird family]." Figure 3 shows an example learning trial. All inference learners were given the nest material verbal predicate and were asked to infer what the bird eats ("What does this bird eat?" along with three possible answers, split into the same two groups of three as in the classification condition and kept throughout learning). The nest material predicate and category label were equally likely to be presented before or after the bird picture (i.e., were equally likely to be the first or third piece of information presented).

All participants received the same four tests in the same order. First, participants completed a novel classification test. For each trial, participants viewed an exemplar and chose one of the 6 family names presented on the screen by clicking on its respective box with the mouse. Next, participants were asked to make confidence judgments about their response ("Likelihood of classification (16%-100%)") using the number pad. There was no time limit on either the classification decision or the confidence judgment. There were 36 trials (6 for each bird family), and they were presented in a random order for each participant.

Next, participants received a verbal predicate classification test. On each trial, participants were presented with one of the studied verbal predicates and were asked to click on which of the 6 bird families it was most likely to be true for. There were 12 verbal predicates (6 food and 6 nest material), and they were presented in a random order for each participant.

Next, participants completed an old/new recognition test with classification. There were three types of exemplars for this test: old exemplars (those presented in the study phase), new exemplars from previously studied categories, and new exemplars from categories never studied in the experiment (these were from other Passeriformes families: Orioles, Jays, Grosbeaks, Vireos, Chickadees, and Thrashers). There were 36 old exemplars, 18 new exemplars from old categories, and 18 new exemplars from new categories. Participants were presented with one bird picture at a time, and the task was to decide whether the presented bird was old or new by clicking the box on the screen that corresponded to their answer. After making the old/new decision, participants classified each bird into one of seven categories. The categories were each of the six families that they had learned about as well as an "Other" category.

Finally, participants received a verbal predicate inference test. For this test, there were two types of questions asked: "What would a [Bird Family] eat?" and "What kind of nest would a [Bird Family] have?" The answers displayed were the six values that corresponded with the question being asked. Both questions were asked about each of the 6 bird families, resulting in 12 trials, which were presented in a random order.

Afterwards, participants completed a questionnaire that asked for each category: "What information did you use to determine if the bird was a_________?" They were also asked to provide any additional information that aided in their decisions. Finally, participants were debriefed and thanked for their time.

Results & Discussion

Learning

The two groups did not differ in learning performance. In the first block of learning, classification learners ($M = 0.75$, $SD = 0.16$) and inference learners ($M = 0.78$, $SD = 0.14$) had similar performance, $t\,(22) < 1$. In the final block of learning, both classification learners ($M = 0.95$, $SD = 0.08$) and inference learners ($M = 0.94$, $SD = 0.09$) performed equally well, $t\,(22) < 1$. Learners in both conditions clearly learned about the verbal features for each category by the end of the learning phase.

Tests

Novel classification. The critical result was performance on novel classification. The verbal information was omitted from the trials and learners had to base their classification decisions on the bird pictures. It was not necessary for either classification or inference learners to pay attention to the bird pictures during the learning phase, so performance above chance would suggest that they went beyond learning what was critical for the learning task to learn what each category was like.

As predicted, inference learners had significantly higher accuracy on the novel classification test ($M = 0.26$, $SD = 0.05$) than classification learners ($M = 0.19$, $SD = 0.08$), t

(22) = 2.63, $p < 0.05$. Inference learners' performance was higher than chance (where chance is 0.167), t (11) = 6.14, $p < 0.001$, while classification learners' performance was not different from chance, t (11) < 1. Inference learners were able to generalize what they had learned about the visual properties common across members of a bird category to novel birds at test while classification learners failed to demonstrate any knowledge of the visual category information.

After each novel classification, participants gave a confidence rating for their response. One participant in the inference condition failed to provide confidence ratings, so is not included in this analysis. Inference learners' confidence ratings, regardless of whether the response was correct or incorrect,[2] (M = 0.34, SD = 0.14) were higher than classification learners' (M = 0.22, SD = 0.07), t (21) = 2.67, $p < 0.05$. Inference learners' higher confidence is most likely because they paid attention to the visual information during learning while classification learners did not; thus they had some idea of what the correct answer could be while classification learners could only guess.[3]

Recognition memory. One explanation for the difference in novel classification performance is that the inherent difference between classification and inference is how learners encode exemplars during learning. Wahlheim, et al. (submitted) suggest that there is a relationship between recognition memory and classification performance, and differences in this relationship across conditions may suggest differences in strategies. While inference learners focus more on what a category is like, perhaps classification learners focus more on individual exemplars. The recognition memory test was included to see if classification learning

[2] The confidence is presented for all responses because the proportion of correct responses is low and different in the two conditions. However, if one looks only at correct responses, inference learners' confidence ratings (M = 0.39, SD = 0.18) were still higher than classification learners' (M = 0.22, SD = 0.06), t (21) = 2.99, $p < 0.01$.

[3] Though this difference is not replicated in Experiment 2, likely due to the addition of catch trials (see Exp. 2 Method).

encouraged an exemplar-based approach to learning about the pictures. If this is the case, then classification learners should have better recognition memory performance (i.e. did you see this exact bird during learning?) than inference learners.[4]

For the recognition memory test, there were three types of trials: Old items (36 items), new items from studied categories (18 items), and new items from unstudied categories (18 items). Table 2 shows recognition memory performance for each type of item by learning condition. There was no advantage for classification learners; performance was slightly better for the inference learners. For only studied category items (both old and new), recognition memory (as measured by hits – false alarms) was similar for inference and classification learners, $t\ (22) = 1.35, p > 0.05$. If items from non-studied categories are included, inference and classification learners still have similar performance, $t\ (22) = 1.21, p > 0.05$. Thus, there is no evidence that either task promotes better recognition memory than the other task.

After participants decided that a bird was old or new, they classified it using one of the six categories of birds or "Other". Table 3 shows classification performance for each type of item by learning condition: old items correctly identified as old in the recognition memory test, old items mistakenly identified as new, new items (broken down by studied categories and unstudied categories) correctly identified as new, and new items mistakenly identified as old. For birds that were old and that participants had correctly classified as old, inference learners were significantly more accurate than classification learners, $t\ (22) = 2.23, p < 0.05$. This is further evidence that inference learners paid more attention to the bird pictures during learning

[4] It is possible that the change from a real-world background to a white background for test could hurt memory for exemplars, but we felt it critical to examine what participants had learned about the specific birds, not just the often distinctive backgrounds. We did consider using the white-background pictures for study as well, but decided it was more important to make the study pictures look as if they were of real-world items photographed in their contexts.

than classification learners. None of the other comparisons showed any clear difference (in all cases, t (22) < 1).

Verbal predicate tests. The verbal predicate classification test and verbal predicate inference test were included to ensure that both groups had learned about the verbal predicates and additionally to see if learners were encoding these features as a function of how they were queried during learning (e.g. do classification learners do better when asked about the verbal predicates as a classification test than as an inference test?). For the verbal predicate classification test (where chance is 0.167), participants had to choose the name of the bird that most likely went with the presented statement. For this test, accuracy was similar for classification learners (M = 0.69, SD = 0.16) and inference learners (M = 0.67, SD = 0.26), t (22) < 1. For the verbal predicate inference test, (where chance is 0.167) participants had to choose the feature value that was most likely true for the bird name presented. For this test, performance was similar for classification learners (M = 0.72, SD = 0.19) and inference learners (M = 0.66, SD = 0.27), t (22) < 1. Both types of learners were able to learn about the verbal predicates and express what they knew equally well in both test formats.

Summary

During the learning phase, we tried to equate as best we could the classification and inference tasks except for the type of information asked about during learning—category label or feature. This inherent difference led to performance differences between the two conditions. The critical result was that inference learners showed better performance than classification learners on a novel *classification* test. This result suggests that inference learners focus on what each category is like, allowing them to better generalize to novel exemplars than classification learners.

CHAPTER 4

EXPERIMENT 2

It is important to examine different means of category learning and understand how they each influence the representation of category knowledge. Previous work contrasting two of the major means of category learning, classification and inference (e.g. Yamauchi & Markman, 1998; Anderson, et al., 2002; Chin-Parker & Ross, 2004; Sakamoto & Love, 2006) has shown distinct differences in what is learned. However, these differences could have been attributed to methodological differences rather than an inherent difference between the two tasks. In Experiment 1, we were able to equate the task with respect to these methodological differences and still found the two learning conditions led to differential performance on the critical novel classification test. Given the novelty and importance of this finding, it is critical to replicate the result. In addition, we attempted to increase both groups' examination of the bird pictures with the introduction of catch trials in which not all the verbal information was presented.

Method

Design

Participants were randomly assigned to one of two between-subjects conditions: classification and inference. There were three counterbalancing variables: Assignment of verbal features to category, answer groupings, and (for inference learners) which of the 2 verbal predicates was inferred (nest material or what the bird eats). These were counterbalanced across participants, resulting in 16 counterbalancing groups.

Participants

There were 32 undergraduate students from the University of Illinois who participated for course credit. One participant was removed from the analysis for having much prior knowledge of bird families.

Materials

The materials were the same as those used in Experiment 1.

Procedure

There were only two minor changes to the Experiment 1 procedure. For the learning phase, we changed the amount of delay between a piece of information being displayed and the next piece of information from 3 s to 5 s. Additionally, for 12 out of 36 trials in each block one of the pieces of verbal information was not displayed (half of the time it was one kind of statement that was omitted and half of the time it was the other kind of statement). For these trials, the picture comprised half of the presented information. We hoped that this would lead to more attention to the pictures across both groups. Everything else was the same as Experiment 1.

Results & Discussion

Learning

As in Experiment 1, the two groups' learning performance did not differ. In the first block of learning, classification learners (M = 0.71, SD = 0.09) and inference learners (M = 0.71, SD = 0.12) had similar performance, t (22) < 1. In the final block of learning, both classification learners (M = 0.92, SD = 0.10) and inference learners (M = 0.90, SD = 0.17) performed equally well, t (22) < 1.

Tests

Novel classification. We replicated the critical result from Experiment 1, that inference learners showed higher novel classification accuracy. The addition of the catch trials did not appear to have much influence on this performance. Inference learners had significantly higher performance on the novel classification test ($M = 0.28$, $SD = 0.09$) than classification learners ($M = 0.19$, $SD = 0.09$), $t\,(29) = 2.55$, $p < 0.05$. Inference learners performed above chance (where chance is 0.167), $t\,(15) = 4.86$, $p < 0.001$, but classification learners did not, $t\,(14) = 1.14$, $p > 0.05$. Inference learners paid more attention to the bird pictures during learning than classification learners and were able to generalize what they had learned about the different birds' visual properties to novel birds.

Inference learners' confidence ratings, regardless of response accuracy,[5] ($M = 0.30$, $SD = 0.11$) were similar to classification learners' ratings ($M = 0.31$, $SD = 0.12$), $t\,(29) < 1$, unlike in Experiment 1 (where inference learners were more confident).[6]

Recognition memory. If classification learners are using an exemplar-based approach to learning, then their recognition memory performance should be better than inference learners' performance. Table 4 shows recognition memory performance for each item type by learning condition. Recognition memory (hits – false alarms), for only studied category items (both old and new), was similar for inference and classification learners, $t\,(29) < 1$. If items from non-studied categories are included, inference and classification learners still show similar

---

[5] If one looks only at correct responses, inference learners' confidence ratings ($M = 0.36$, $SD = 0.16$) were similar to classification learners' ($M = 0.36$, $SD = 0.13$), $t\,(29) < 1$.

[6] The addition of catch trials may have increased classification learners' attention to the bird pictures during learning, thus increasing their confidence and making it similar to inference learners'.

performance, t (22) < 1. There is no evidence that one learning task encourages an exemplar-based encoding strategy more than the other[7].

After participants decided that a bird was old or new, they classified it using one of the six categories of birds or "Other". Table 5 shows classification performance for each type of item by learning condition: Old items correctly identified as old, old items mistakenly identified as new, new items (broken down by studied and unstudied categories) correctly identified as new, and new items mistakenly identified as old. There were no differences between classification and inference learners' performance (t (29) < 1, except performance for birds that were new (from unstudied categories) that participants had incorrectly classified as old, t (29) = 1.19, $p > 0.05$).

Verbal predicate tests. As in Experiment 1, these tests were included to ensure that participants had learned the verbal predicates as well as to see if learners encoded the verbal predicates differently based on learning task. Both types of learners performed equally well on both tests (where chance is 0.167 for both tests). For the verbal predicate classification test, performance was similar for classification learners (M = 0.81, SD = 0.26) and inference learners (M = 0.79, SD = 0.23), t (29) < 1. For the verbal predicate inference test, performance was similar for classification learners (M = 0.84, SD = 0.25) and inference learners (M = 0.82, SD = 0.22), t (29) < 1.

[7] As in Experiment 1, items used for the recognition test used a white background instead of a real-world background to reduce reliance on context and examine what participants learned about the specific birds. It is possible that this change could have affected exemplar-based memory for the items.

Summary

We successfully replicated the main result from Experiment 1. Inference learners performed significantly above chance on the novel classification test, whereas the classification learners did not. This result suggests that differences in processing lead to differences in the kinds of knowledge acquired during learning.

CHAPTER 5

GENERAL DISCUSSION

Categories are useful for a broad range of tasks, yet the relationship between use and learning has not been widely studied. Classification and inference are essential category uses that can be considered formally equivalent, yet there are clear differences in what is learned. Our experiments were designed to address whether differences between classification and inference learning could be explained by methodological problems with equating the two tasks or were due to inherent task differences. We successfully equated for three differences between these tasks that were previously unaddressed in the literature: number of queried dimensions, exception features, and type of queried feature. The only differences between the two learning tasks were the type of information queried (label or feature) and whether the provided verbal predicate was a feature or a label. Even when controlling for methodological problems, we found performance differences for classification and inference learners. Inference learners outperformed classification learners on the novel classification test, suggesting that these two tasks lead to very distinct strategies, and consequently learners attend to different information about the categories. In the remainder of the General Discussion we will address the nature of this difference, both inside and outside of the laboratory, and implications for category learning and use beyond classification and inference as they have been studied in the past.

Classification and Inference Learning: Similarities and Differences

Classification and inference are similar in many ways. In both tasks, learners are providing a piece of missing information based on other available information. If we take the view that a category label is a feature (Anderson, 1991), then classification and inference have the same goal. The goal in both cases is to find the information that best predicts the dimension being inferred, whether the dimension is a label or a feature.

However, category labels may not be psychologically equivalent to features (c.f. Markman & Ross, 2003). Category labels are arbitrary names referring to classes of objects, while features are meaningful parts of an object, and there is evidence that people do treat them differently (e.g. Gelman & Heyman, 1999). Because an item's features relate in meaningful ways to the feature in question, whereas category labels are not meaningfully related to the item's features, inference learners pay more attention to the item's other features than classification learners do. Thus the critical difference between classification and inference, what is being inferred, can be explained by this psychological distinction between a category label and a feature. Here we more closely examine how inferring a label and inferring a feature may lead learners to attend to different kinds of information.

When classifying an item (i.e. inferring a category label), one needs to contrast all possible category labels based on the presented features. The goal is to determine that the object is in one category as opposed to other known categories, and the best way to accomplish the goal is to learn the information that differs between categories and therefore best predicts category membership (Yamauchi & Markman, 1998; Chin-Parker & Ross, 2004). To determine what kind of dog is in front of you, the most helpful information will be its diagnostic features (e.g. whether it has pointy ears, its coloring, and its build), and making this classification may not be improved by knowing feature relations or less predictive properties of the category.

Alternatively, when inferring a feature of an object from an already specified category, one needs only to think about that one category and figure out how the combination of presented features correlates with each of the inferred feature's values. The goal is to determine the value of the feature that fits best with the other presented information, and the best way to accomplish the goal is to learn the relations between the feature being inferred, the presented features, and the category label (Chin-Parker & Ross, 2002). Deciding if a dog will bite or not is not just a matter of knowing what kind of dog it is, but also knowing about the likelihood of that kind of dog biting based on other features that are present.

Classification and Inference Outside the Laboratory

It was necessary to show that classification and inference lead to differences in performance when everything is controlled to demonstrate that the differences lie in the tasks themselves and not in the conditions under which they are performed. In their real world forms, however, these two tasks are very different from each other in a variety of ways. Here we provide examples for each task to demonstrate that both are implemented differently, and in more complex ways, outside the laboratory and to speculate on the implications. We focus on the variety of queries, the interrelatedness of features in real-world items, the knowledge accessed, and the influence of uncertainty.

First, although we queried only one feature in the inference condition for methodological purposes, inferences are usually made about a variety of properties, while classifications, by definition, require one to determine the category label. One cannot easily predict which properties of objects will need to be inferred, and this uncertainty would likely lead to attention to all of the various features of a category (e.g., Rehder et al., 2009). Queried features are learned about more than non-queried (e.g., Anderson et al., 2002), so querying multiple features for inference is likely to lead to good learning about a variety of features.

Second, the interrelatedness of features in real-world items differentially influences the two tasks. In this paper, and in other work contrasting classification and inference using artificial stimuli, the values of the experimenter-controlled features (in this case the verbal features, e.g. Finches eat nuts) were arbitrarily assigned with respect to the category and the other features. For real world categories, a category's features are usually correlated with other features. For example, inferring whether or not a dog will bite will be correlated with whether or not the dog has bared its teeth and whether or not it is growling. Inference learners are more likely to learn about these within-category correlations than classification learners (Chin-Parker & Ross, 2002). Knowing that these correlations usually exist across category members could be one reason why inference learners are more likely to learn about what a category is like beyond what is necessary for the task. In other words, inference learners actively look for correlations between features because they expect them to be there and they may be useful for making a future inference. The querying of multiple features (previous paragraph) may increase the likelihood of noticing these inter-feature correlations in the inference task.

Third, the knowledge accessed is very different in the two tasks and likely to have implications for what is learned. In the typical classification-learning paradigm there are only two possible labels (though here there were 6 categories), so the knowledge required is very limited. When classifying an object in a real world situation, however, there are often many contrasting categories whose knowledge needs to be accessed and considered before determining the appropriate label. Thus, much of the knowledge considered is of other categories and the goal is to learn which category to focus on. Inference learners, however, are given the category and can access knowledge about that single category and its features and relations. Thus, all the focus in inference learning is on a known category and what features a member of that category is likely to have (given its other features). When there are only two (or some small number of) categories to contrast, as in typical laboratory settings, this amounts to a small difference between these two tasks. In the real world, however, when there are many different categories to consider, this may be a critical difference between the tasks.

Fourth, uncertainty may influence both tasks, though differently. A common uncertainty when classifying objects in the real world is that not all of the object's features are known. There may be features not evident (e.g., internal organs, food preferences) or features not visible because three-dimensional objects cannot be seen in their entirety. In contrast, in the typical laboratory classification-learning task, all the features of objects are presented on the screen. Taylor and Ross (2009) presented only some of the features on each trial and found that classification learning now led to gaining more knowledge about other common but non-predictive features of the categories. Uncertainty for inference learners could be of various kinds. For example, one type of uncertainty that has been examined is when the category label is ambiguous, such as when one gets a brief glimpse of an object. In these cases, people often treat the most likely category as if it certain (e.g. Murphy & Ross 1994; 2007; though see Hayes & Newell, 2009). The main idea of this section is that the experimental implementations that have tried to contrast these two tasks may reduce common differences between them that have important implications for their use in more real-world settings.

Category Learning and Use

There are many tasks that require the use of categories, and what is learned is likely to be related to how categories are used. One important distinction in category use is between determining what category something is in, classification, and using category knowledge for some purpose, such as inference, problem solving, understanding, or constructing explanations. What is learned from classification is most likely different from what is learned in each of these other tasks.

Understanding how these tasks separately affect learning is important, but we typically classify in order to perform one of these other tasks. Furthermore, one commonly uses the same category in numerous ways. The resulting category representation is a reflection of many types of interactions with category members (Markman & Ross, 2003). Now that we have a better idea of how these tasks, specifically classification and inference, separately affect what is learned, we can focus on what happens when they interact during learning.

Not only does the kind of task affect what is learned about a category, so does the way in which the task is implemented. For instance, inferences can be made about different types of properties, diverse category structures and formats, and in unique contexts. Attention to specific information in an inference task may be modulated by these variables. A complete understanding of category learning and use requires a more detailed explanation about how these tasks impact what is learned under a variety of circumstances.

TABLES AND FIGURES

Table 1

Family resemblance category structure. Each dimension is a feature (e.g. wing type) and each feature has two values (e.g. striped wings and spotted wings), indicated by 1 and 0.

		Feature dimensions			
	Exemplars	1	2	3	4
	Prototype	1	1	1	1
	1	1	1	1	0
Category A	2	1	1	0	1
	3	1	0	1	1
	4	0	1	1	1
	Prototype	0	0	0	0
	1	0	0	0	1
Category B	2	0	0	1	0
	3	0	1	0	0
	4	1	0	0	0

Table 2

Experiment 1 recognition memory performance. Means and standard deviations by item type and condition.

	Item Type					
	Old/Studied Categories		New/Studied Categories		New/Not Studied Categories	
	Hits	Misses	Correct Rejections	False Alarms	Correct Rejections	False Alarms
Classification						
Mean	0.59	0.41	0.58	0.42	0.70	0.30
SD	0.22		0.22		0.21	
Inference						
Mean	0.54	0.46	0.72	0.28	0.82	0.18
SD	0.21		0.20		0.14	

Table 3

Experiment 1 post-recognition classification accuracy. Means and standard deviations by item type, participants' recognition memory response, and condition.

	Item Type					
	Old/Studied Categories		New/Studied Categories		New/Not Studied Categories	
Participant Response:	Old	New	Old	New	Old	New
Classification						
Mean	0.20	0.08	0.27	0.08	0.06*	0.74*
SD	0.11	0.11	0.17	0.10	0.15	0.35
# Observations	255	177	90	126	64	152
Inference						
Mean	0.31	0.10	0.27	0.08	0.05*	0.71*
SD	0.15	0.06	0.19	0.10	0.09	0.23
# Observations	234	198	60	156	39	177

*Proportion of trials correctly classified as "Other"

Table 4

Experiment 2 recognition memory performance. Means and standard deviations by item type and condition.

	Item Type					
	Old/Studied Categories		New/Studied Categories		New/Not Studied Categories	
	Hits	Misses	Correct Rejections	False Alarms	Correct Rejections	False Alarms
Classification						
Mean	0.59	0.41	0.61	0.39	0.74	0.26
SD	0.16		0.19		0.17	
Inference						
Mean	0.63	0.37	0.56	0.44	0.75	0.25
SD	0.14		0.20		0.17	

Table 5

Experiment 2 post-recognition classification accuracy. Means and standard deviations by item type, participants' recognition memory response, and condition.

	Item Type					
	Old/Studied Categories		New/Studied Categories		New/Not Studied Categories	
Participant Response:	Old	New	Old	New	Old	New
Classification						
Mean	0.21	0.13	0.20	0.10	0.06*	0.51*
SD	0.10	0.13	0.26	0.08	0.16	0.30
# Observations	316	224	105	165	69	201
Inference						
Mean	0.13	0.23	0.25	0.10	0.04*	0.64*
SD	0.14	0.12	0.20	0.09	0.07	0.31
# Observations	361	215	126	162	72	216

*Proportion of trials correctly classified as "Other"

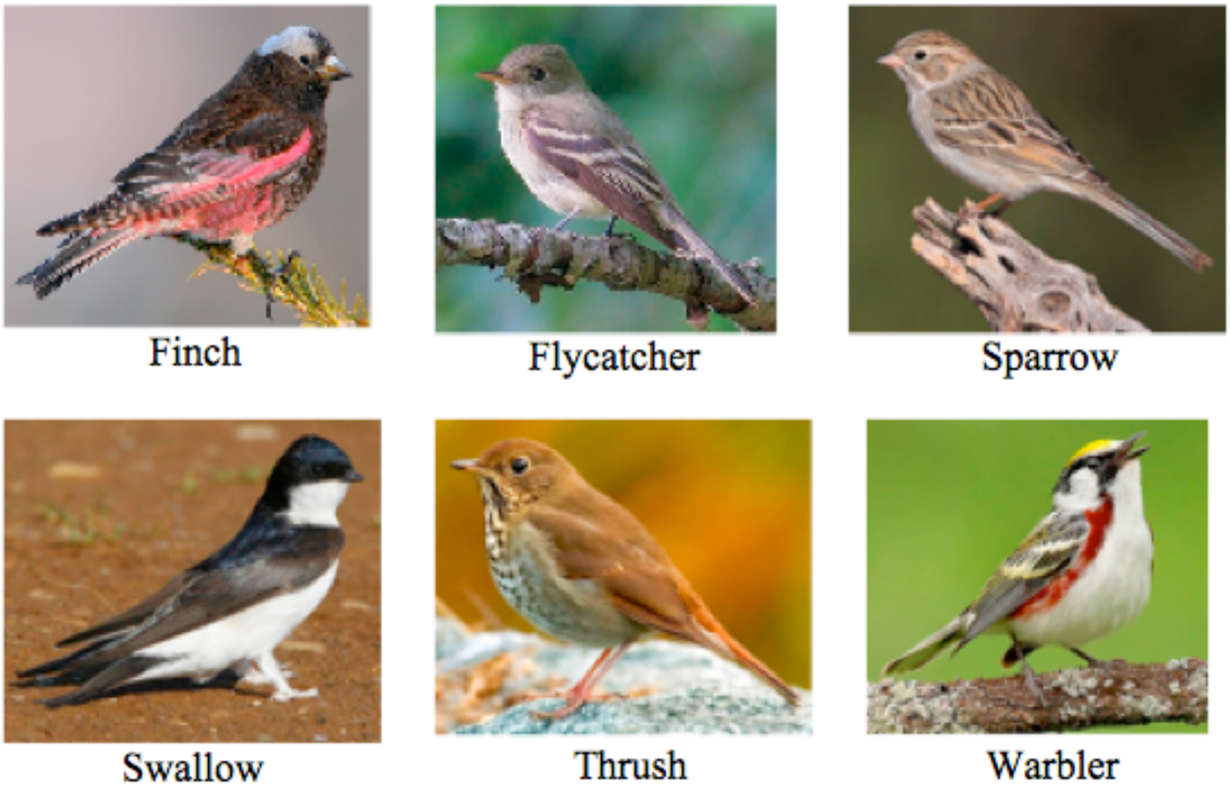

Figure 1. Example bird stimuli.

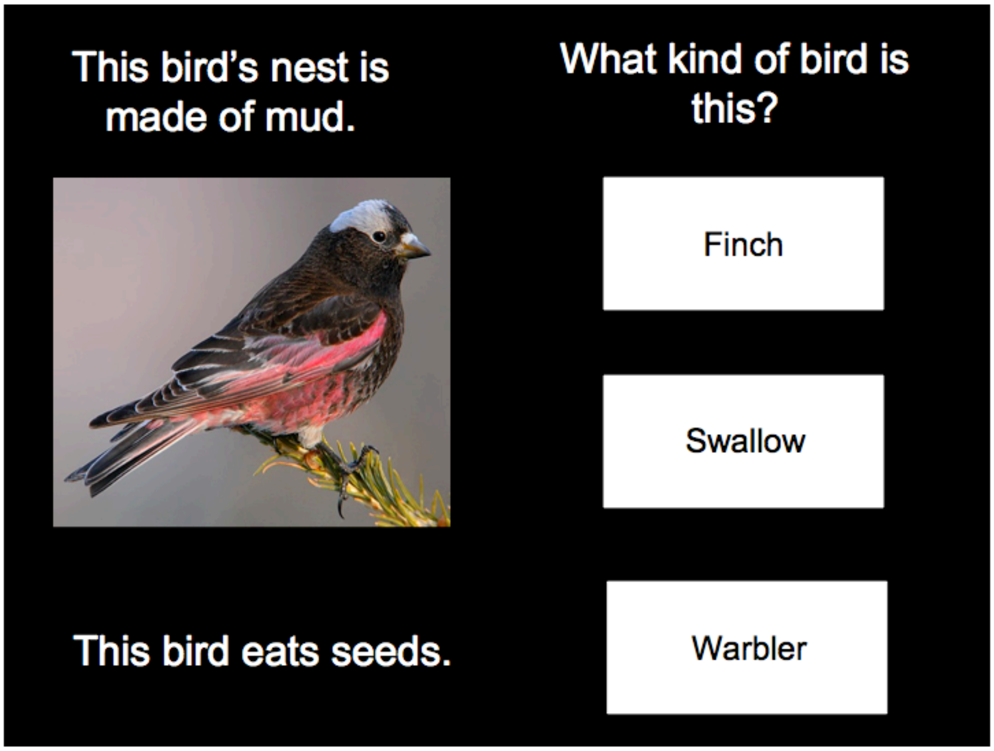

Figure 2. Example classification learning trial. Information appeared on the screen one piece at a time until everything was displayed. The verbal predicates were equally likely to appear in either location on the screen.

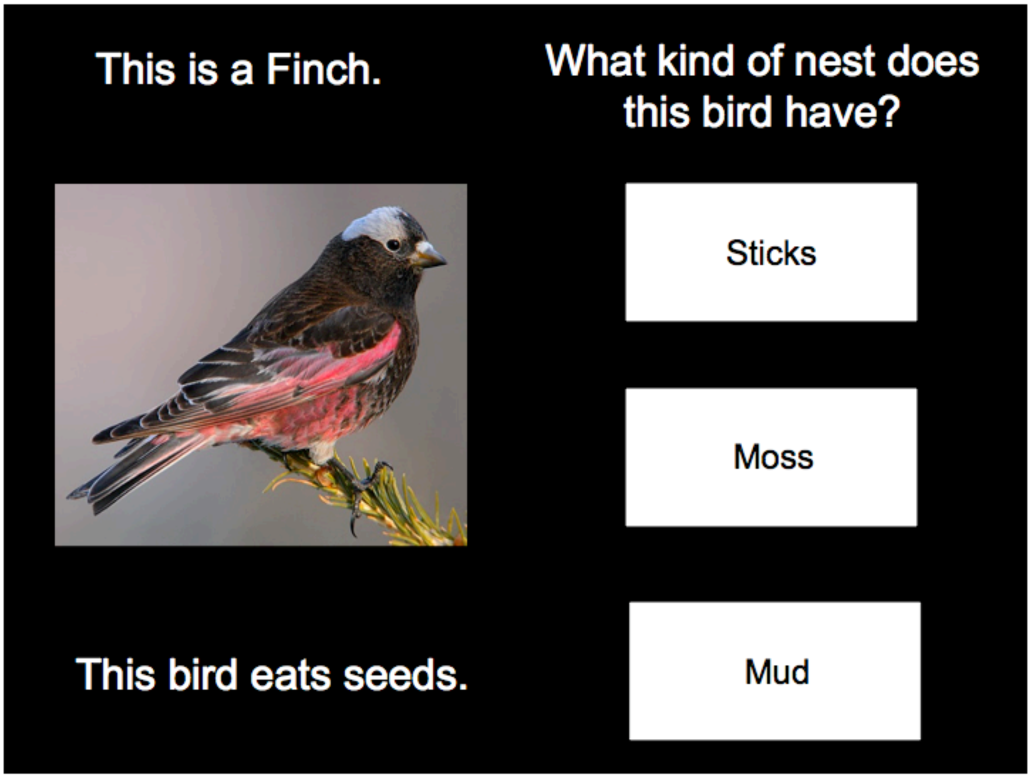

Figure 3. Example inference learning trial. Information appeared on the screen one piece at a time until everything was displayed. The verbal predicates were equally likely to appear in either location on the screen.

REFERENCES

Anderson, A.L., Ross, B.H., & Chin-Parker S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30,* 119-128.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409-429.

Chin-Parker, S. & Ross, B.H. (2002). The effect of category learning on sensitivity to within category correlations. *Memory & Cognition, 30,* 353-362.

Chin-Parker, S. & Ross, B.H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 216-226.

Gelman, S.A., & Heyman, G.D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10*, 489-493.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition, 37,* 730-743.

Johansen, M. K. & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1433-1458.

Markman, A.B., & Ross, B.H. (2003). Category use and category learning. *Psychological Bulletin, 129*, 592–613.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Murphy, G. L. & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27,* 1148-1193.

Murphy, G. L., & Ross, B. H. (2007). Use of single or multiple categories in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 205-225). Cambridge: Cambridge University Press.

Rehder, B., Colner, R. M., & Hoffman, A. B. (2009). Feature inference learning and eyetracking. *Journal of Memory and Language, 60,* 394-419.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 1573–1605.

Sakamoto, Y. & Love, B. C. (2006). Sizeable sharks swim swiftly: Learning correlations through inference in a classroom setting. In R. Sun & N. Miyake (Eds.), *Proceeedings of the 28th Annual Conference of the Cognitive Science Society.*

Wahlheim, C. N., Jacoby, L. L., & Coane, J. H. (submitted). Retrieval practice enhances natural category learning: Effects on accuracy and on metacognition at the levels of items and categories.

Yamauchi, T., & Markman, A.B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124–148.